Acc homoeoloci and the evolution of wheat genomes

D. Chalupska*, H. Y. Lee*, J. D. Faris†, A. Evrard‡, B. Chalhoub‡, R. Haselkorn*§, and P. Gornicki*§

*Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, IL 60637; †United States Department of Agriculture, Agricultural Research Service, Cereal Crops Research Unit, Northern Crop Science Laboratory, Fargo, ND 58105; and ‡Unité de Recherche en Génomique Végétale, Institut National de la Recherche Agronomique, Organization and Evolution of Plant Genomes, 2 Rue Gaston Crémieux, F91150 Evry, France

Contributed by R. Haselkorn, April 25, 2008 (sent for review January 30, 2008)

The DNA sequences of wheat *Acc-1* and *Acc-2* loci, encoding the plastid and cytosolic forms of the enzyme acetyl-CoA carboxylase, were analyzed with a view to understanding the evolution of these genes and the origin of the three genomes in modern hexaploid wheat. *Acc-1* and *Acc-2* loci from each of the wheats *Triticum urartu* (A genome), *Aegilops tauschii* (D genome), *Triticum turgidum* (AB genome), and *Triticum aestivum* (ABD genome), as well as two *Acc-2*-related pseudogenes from *T. urartu* were sequenced. The 2.3–2.4 Mya divergence time calculated here for the three homoeologous chromosomes, on the basis of coding and intron sequences of the *Acc-1* genes, is at the low end of other estimates. Our clock was calibrated by using 60 Mya for the divergence between wheat and maize. On the same time scale, wheat and barley diverged 11.6 Mya, based on sequences of *Acc* and other genes. The regions flanking the *Acc* genes are not conserved among the A, B, and D genomes. They are conserved when comparing homoeologous genomes of diploid, tetraploid, and hexaploid wheats. Substitution rates in intergenic regions consisting primarily of repetitive sequences vary substantially along the loci and on average are 3.5-fold higher than the *Acc* intron substitution rates. The composition of the *Acc* homoeoloci suggests haplotype divergence exceeding in some cases 0.5 Mya. Such variation might result in a significant overestimate of the time since tetraploid wheat formation, which occurred no more than 0.5 Mya.

acetyl-CoA carboxylase | *Triticeae* | grass

Interest in the genes encoding acetyl-CoA carboxylase (ACCase) stems from the central importance of the enzyme in lipid metabolism. In the plastid, ACCase encoded by the *Acc-1* gene catalyzes the first step in fatty acid synthesis, converting acetyl-CoA to malonyl-CoA, where it is used to add two-carbon fragments to the growing fatty acid chain. In the cytoplasm, ACCase encoded by the *Acc-2* gene makes malonyl-CoA for secondary metabolism including the extension of long-chain fatty acids. The gene encoding plastid ACCase in grasses, which replaced the multisubunit plastid enzyme of bacterial origin still found in dicots, arose by duplication of a gene encoding the multidomain cytosolic ACCase (1). One copy of the ancestral gene acquired a plastid-targeting signal (1, 2) and a new promoter and new regulatory elements (2, 3). The wheat *Acc-1* and *Acc-2* genes show no sequence similarity outside the exons encoding the mature enzyme (2, 3). For both *Acc-1* and *Acc-2*, all three *Triticum aestivum* homoeologs are transcriptionally active, and each uses two nested promoters and alternative splicing of the first intron to produce transcripts with different organ specificity. Each of the homoeologs contributes equally to the total mRNA level of each transcript type (2, 3).

Analysis of genomic DNA provided evidence of a single copy of an active *Acc* gene in all of the homoeologous genomes of the hexaploid wheat *T. aestivum* (1, 4–8). One *Acc-1* gene and one *Acc-2* gene are present in rice. Sampling of wheat relatives (1, 4, 5) identified two copies of *Acc-1* in diploid *Aegilops speltoides.* Topology of the *Acc-1* phylogenetic tree (in figure 1 of ref. 4) suggests that the gene was duplicated at least once before divergence of the *Aegilops* and *Triticum* genomes, and this was followed by the loss of a different copy of the gene in different lineages. *Acc-1* genes are located on the short arm of wheat group 2 chromosomes. The *Acc-2* gene was also duplicated at least twice in *Triticeae* evolution followed by specific gene loss in some lineages (5). An *Acc-2*-related partially processed pseudogene has been found in all *Triticeae* species tested (1). *Acc-2* genes are located on the long arm of group 3 chromosomes (5, 6).

The wheat *Acc* genes, because of their large size and the availability of functional information, provide an excellent starting point for comparative genome analysis. The transcribed part of the wheat *Acc-1* gene is ≈13.5 kb with ≈6.9 kb of coding sequence (2,311 amino acids including the putative chloroplast transit peptide) and has 33 introns with a total length of ≈6.3 kb. The transcribed part of the *Acc-2* gene is ≈12.3 kb with ≈6.8 kb of coding sequence (2,258 or 2,260 amino acids) and has 29 introns with total length of ≈4.4 kb.

Evolution of *Triticum* and *Aegilops* has been the focus of studies driven by interest in wheat as a major crop and by the complexity of their evolutionary history: from the divergence of wheat diploid ancestors, through two steps of polyploidization, domestication and extensive breeding. In wheat, most active genes are highly dispersed between long rapidly changing intergenic regions consisting primarily of repetitive sequences. In addition to mutations, active genes are affected by frequent gene duplication and loss, as well as conversion and unequal crossing over between paralogs leading to hybrid genes. These processes are lineage-, genome-, and gene copy-specific. In addition to the recent polyploidization in wheat, earlier segmental duplication and whole genome duplication, such as the ancient duplication predating divergence of the major grass subfamilies, and chromosome rearrangements and segmental loss (9–11), add complexity to the comparative genome analysis.

The time of formation of the hexaploid wheat genome, *T. aestivum* (ABD), was estimated to be 10 thousand years ago (12), based on archeological artifacts. The divergence times of the wheat diploid ancestors, *Triticum urartu* (A genome), *Aegilops tauschii* (D genome), and *A. speltoides* (S genome, a progenitor of the B genome), of 2–4 Mya and the time of formation of tetraploid wheat, *Triticum turgidum* (AB genome) of <0.5 Mya, were previously estimated by using the molecular clock method (4, 13). An earlier divergence of the progenitors of the A, B, and D genomes was proposed based on analysis of the genes *Sut1* and *PhyC* (14). All these estimates carry the possible error associated with dating fossils, which was used to place the divergence of the major grass subfamilies. Our studies on wheat evolution calibrate the clock by using 60 Mya for the divergence of *Pooideae* and *Panicoideae*. The separation of the major clades of grasses has been estimated at 55–83 Mya (15–17), whereas an unpublished study of six fossils yielded a date of 52 ± 8 Mya (E. Kellogg, personal communication).

Author contributions: J.D.F., A.E., B.C., R.H., and P.G. designed research; D.C., H.Y.L., J.D.F., A.E., B.C., and P.G. performed research; A.E., B.C., and P.G. contributed new reagents/analytic tools; J.D.F., A.E., B.C., R.H., and P.G. analyzed data; and J.D.F., R.H., and P.G. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU626553, EU626554, and EU660891–660903).

§To whom correspondence may be addressed. E-mail: rh01@uchicago.edu or pg13@uchicago.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803981105/DCSupplemental.

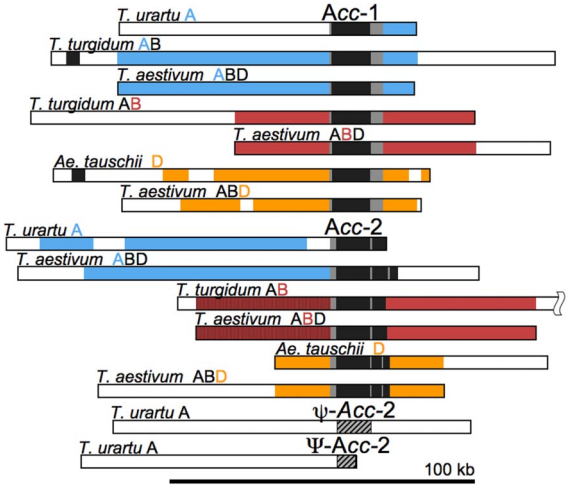

Fig. 1. Sequenced *Acc* loci. The coding part of *Acc* genes and conserved genes located upstream of *Acc-1* and immediately downstream of *Acc-2* are shown in black. Pseudogenes are indicated by black stripes. Sequence similarity between *Acc-1* and *Acc-2* loci is limited to coding exon sequences. Regions flanking the *Acc* coding part shown in gray contain sequence segments conserved between the A, B, and D homoeologs. Sequence similarity among homoeologs is indicated by colors. Results of pairwise sequence comparisons are shown in Fig. S1. The darker color for the segment upstream of the B *Acc-2* gene indicates a lower level of sequence identity than in the downstream segment. The white segments indicate lack of orthologous sequence similarity. Annotation is available at www.tigr.org/tdb/e2k1/tae1.

In an earlier work, we studied *Acc* evolution in hexaploid wheat and a broad sample of its relatives, based on the sequences of a gene fragment (1, 4, 5). Here, we analyze full-length *Acc* genes and their genomic context in diploid, tetraploid, and hexaploid wheat to extend our understanding of the evolution of these genes and the wheat genomes.

Results

BACs containing all of the *Acc* genes predicted on the basis of previously published sequence information (1, 4, 5) were isolated [Fig. 1 and supporting information (SI) Table S1], except for the A homoeolog of *Acc-2* from *T. turgidum*. In the latter case, all three *Acc-2* BACs isolated from the *T. turgidum* library contained only the B homoeolog. The *Acc-2,2* gene was found previously only in a wheat genomic DNA library (5, 7); it could not be confirmed in any of our follow-up experiments, including the BAC analysis described in this paper. *T. turgidum* and *T. urartu* BACs containing an *Acc-2*-related partially processed pseudogene (*ψ-Acc-2*) were also found. The available *T. urartu ψ-Acc-2* sequence is not complete, but it contains sequences corresponding to introns 2 and 3 of the active *Acc-2* gene, one premature stop codon and one frame shift (Fig. 2). This pseudogene was previously found to be present in all homoeologous chromosomes of wheat and in its diploid relatives, including rye and barley (1). A second copy of the *Acc-2* gene [previously called *Acc-2* II (1)] was found in *T. urartu* BACs. This gene is a pseudogene (*Ψ-Acc-2*). It contains sequences corresponding to the coding exons and 28 introns found in active *Acc-2* genes, but it has two in-frame stop codons and two frame shifts (Fig. 2). The frame shifts were confirmed by PCR amplification and sequencing of the corresponding fragment of *T. urartu* genomic DNA. It appears to be a "young" pseudogene, as suggested by a high ratio of synonymous to nonsynonymous substitutions. It acquired some of the deleterious mutations only very recently; the two frame shifts found in *T. urartu* are not present in *Triticum monococcum* or *T. turgidum* (determined by PCR amplification and sequencing of a fragment of genomic DNA). This gene was previously found, based on partial sequence, to be present in *T. urartu*, *T. monococcum*, and

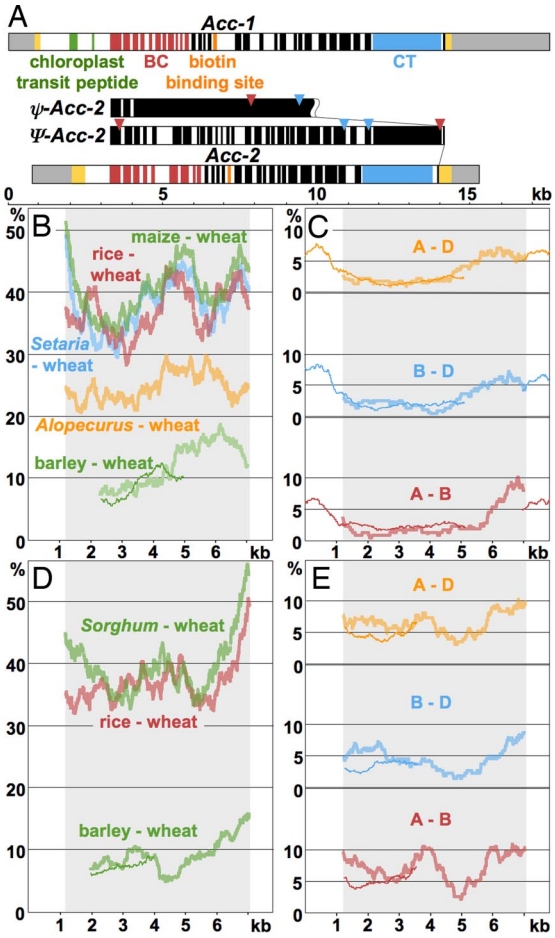

Fig. 2. Structure of *Acc* genes and pairwise sequence comparisons at synonymous and intron sites. (*A*) Exons encoding ACCase highly conserved BC and CT domains are shown in the same color. Introns are shown in white. 5′- and 3′-UTRs are shown in yellow. 5′- and 3′-ends of the genes shown in gray contain sequence segments conserved between A, B, and D homoeologs of the *Acc-1* and *Acc-2* genes. There is no sequence similarity between *Acc-1* and *Acc-2* outside the coding exons. Red and blue triangles indicate positions of premature stop codons and frame shifts, respectively. The diagram depicts genes from *T. urartu*, but the structure of *Acc* genes from other wheat species is very similar. (*B–E*) Pairwise sequence comparisons of *Acc-1* (*B* and *C*) and *Acc-2* (*D* and *E*) genes at synonymous sites (thick line) and at concatenated noncoding and intron sites (thin line). The percentage of nucleotide difference was calculated by using a 1-kb window sliding in 1 bp. The first nucleotide of the first intron in the coding region (highlighted in gray) is lined up with the first nucleotide of the translation start codon. The intron and exon sequences do not line up precisely because of their varying position and length. The last intron was concatenated with the 3′-end sequence of the gene. The 5′- and 3′-end sequences of the *Acc-1* gene used in these comparisons include only large conserved segments. Only partial *Acc* sequences from barley were available.

Triticum armeniacum (A and presumably G homoeolog); in *Triticum timopheevii* (presumably G homoeolog); and in *T. turgidum* (presumably B homoeolog) (5) but not in any other species tested by PCR cloning/sequencing or by screening genomic and BAC libraries (ref. 5 and this paper). Most notably, it appears to be absent in *T. aestivum* and in *A. tauschii*.

None of the sequenced loci contains more than one *Acc* gene or pseudogene. *Acc-2* genes and *Acc-2*-related pseudogenes are not within a short distance of each other (same BAC insert). However, they are located in the same chromosome segment, as shown by chromosome-mapping experiments (5). This observation is also consistent with different chromosomal localizations of *Acc-1* and *Acc-2* genes. The structure of *Acc* genes is shown in Fig. 2*A*. The

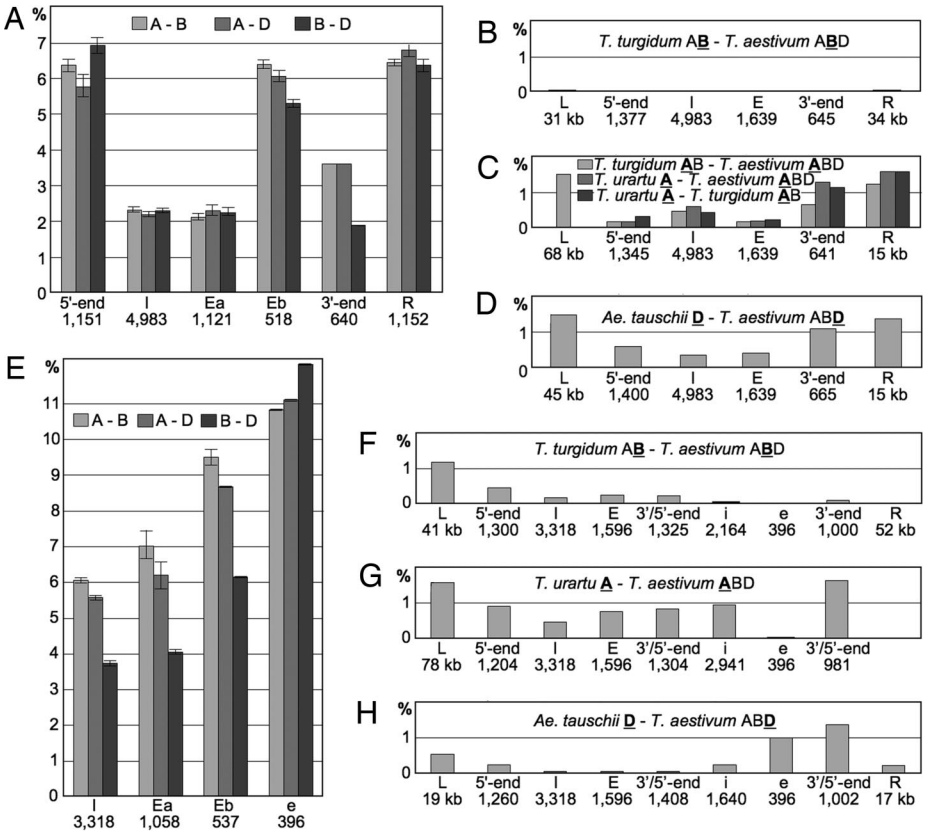

Fig. 3. Nucleotide substitution rates in different parts of the *Acc-1* (*A–D*) and *Acc-2* (*E–H*) loci. Synonymous substitution rates are shown for the coding sequences: Ea, from translation start codon to the beginning of the CT domain; Eb, CT domain to the stop codon; E, full-length coding sequence. Intron sequences (I) include all introns, except the first (leader) intron, which was concatenated at the end of the 5′ -end segment, and the last intron, which was concatenated at the beginning of the 3′-end segment. 3′/5′-end indicates the entire region between the stop codon and the start codon of *Acc-2* and the genes found immediately downstream of *Acc-2* (Fig. 1). i and e, intron and exon sequences, respectively, of the gene immediately downstream of *Acc-2*. Only large conserved blocks of sequences upstream and downstream of the coding part of the *Acc-1* gene were used in these comparisons. The number of nucleotides or number of synonymous sites in exons is shown for each segment. Error bars in *A* and *E* indicate standard deviations for all pairwise distances. The underlined letter in each species genome composition indicates compared homoeoloci.

distribution of synonymous and intron substitutions along the *Acc* genes and nucleotide substitutions in sequences flanking the *Acc* genes is shown in Fig. 2 *B–E* and Fig. S1. Intron sequence divergence does not allow any comparisons above the *Triticeae* tribe level, and sequence divergence in the flanking regions does not allow any comparisons other than among homoeologs. The distribution of *Acc-1* synonymous substitutions between species representing the three grass subfamilies wheat and *Alopecurus* (*Pooideae*), rice (*Ehrhartoideae*) and maize, and *Sorghum* and *Setaria* (*Panicoideae*) (Fig. 2*B*) varies significantly along the coding part of *Acc* (Fig. 2) in a pattern suggesting correlation with sequence conservation at the amino acid level; the biotin carboxylase (BC) and carboxyltransferase (CT) domains are the most conserved parts of ACCase. Their coding sequences have lower incidence of nonsynonymous substitution but also lower incidence of substitutions at synonymous sites. The pattern is less profound for *Acc-2* (Fig. 2*D*). Note that this gene does not encode a transit peptide, which is less conserved. As a consequence, the conserved BC domain is located very close to the 5′ terminus of the gene.

Acc-1 comparisons for barley and wheat and between wheat A, B, and D homoeologs show an increased incidence of synonymous substitutions near the 3′-end of the gene (Fig. 2 *B* and *C*). For the first two-thirds of the ACCase gene (from the translation start to the beginning of the CT domain) containing all but two introns (Fig. 2*A*), the synonymous and intron substitution rates are very similar. The genetic distances between barley and wheat increase with distance from the translation start codon, similar to the increase of the synonymous site distance between wheat and *Alopecurus*, *Setaria*, rice, and maize (Fig. 2*B*). The distances along this part of the gene remain similar for all three pairs of homoeologs (Fig. 2*C*). The increased synonymous substitution rate in the last one-third of the gene (CT domain) is due to accumulation of homoeolog-specific substitutions. This phenomenon is most obvious for the *Acc-1* B homoeolog where 14 of 36 mutations (40%) are found in a stretch of 418 nucleotides (6%). These homoeolog-specific substitutions also contribute to the observed higher genetic distance between 3′-end fragments of the gene when wheat and barley sequences are compared.

For the first two-thirds of the *Acc-1* gene (segment Ea and I, Fig. 3*A*), the synonymous substitution rate and the intron substitution rate are equal for the three pairs of homoeologs. The synonymous substitution rate in the last third of *Acc-1* (segment Eb Fig. 3*A*) is three times higher, reflecting the accumulation of synonymous substitutions near the end of the gene (above). The substitution rates are also 3-fold higher in the conserved segments at the 5′-end of the gene (including putative promoters and regulatory elements, 5′-UTR, and the first intron) and immediately downstream of the gene (R segments, Fig. 3*A*). The rate remains the same or is only 50% higher in the 3′-end segment including the last intron, 3′-UTR, and short stretch of highly conserved sequence following the transcription termination site (Fig. 3*A*).

Very few nucleotide substitutions were found in the 70-kb fragment of the *Acc-1* B locus in *T. turgidum* and *T. aestivum* (Fig. 3*B* and Fig. S1*A*), consistent with the formation of the ABD hexaploid from an AB tetraploid as recently as ten thousand years ago. The distances, however, are much higher among the A and D *Acc-1* homoeologs (Fig. 3 *C* and *D*). This difference may be explained by haplotype variation and, for the *T. urartu* and *T. turgidum* pair, by the age of the AB tetraploid. The nucleotide substitution rates upstream and downstream of the *Acc-1* gene are significantly higher (2- to 8-fold) than either intron or synonymous rates within the gene (L and R segments, Fig. 3 *C* and *D*). These rates vary significantly across the locus (Fig. S1*A*).

For the first two-thirds of *Acc-2* (segment Ea and I, Fig. 3*E*), the synonymous substitution rate and the intron substitution rate are equal for the three pairs of homoeologs. The synonymous substitution rate is two times higher in the last third of the *Acc-2* gene and in the gene located immediately downstream (segment Eb and e,

EVOLUTION

Table 1. Pairwise divergence times of grass genomes based on ACC genes

Genome pairs	Full-length ACC		Ea fragment of ACC		All but two introns	
	Number of substitutions per synonymous site	Divergence time, Mya	Number of substitutions per synonymous site	Divergence time, Mya	Number of substitutions	Divergence time, Mya
Acc-1						
Pooideae– Panicoideae	0.576 ± 0.029	60	0.577 ± 0.036	60	n.a.	n.a.
Rice–*Panicoideae*	0.559 ± 0.028	60	0.556 ± 0.031	60	n.a.	n.a.
Rice–*Pooideae*	0.502 ± 0.023	52.2 ± 2.4	0.494 ± 0.030	51.4 ± 3.1	n. a	n.a.
Maize–*Setaria*	0.261 ± 0.015	27.2 ± 1.5	0.251 ± 0.018	26.1 ± 1.8	n.a.	n.a.
Wheat–*Alopecurus*	0.299 ± 0.016	31.2 ± 1.7	0.292 ± 0.021	30.4 ± 2.2	n.a.	n.a.
Wheat–*Brachypodium*	0.312 ± 0.017	32.4 ± 1.8	0.316 ± 0.020	32.9 ± 2.1	n.a.	n.a.
Wheat–barley	0.125 ± 0.011 (157)	13.6 ± 1.2	0.107 ± 0.011 (84)	11.8 ± 1.8	0.092 ± 0.005 (317)	11.8
A–B	0.033 ± 0.004 (53)	3.5 ± 0.5	0.019 ± 0.004 (21)	2.0 ± 0.5	0.022 ± 0.002 (78)	2.8 ± 0.3
A–D	0.034 ± 0.005 (54)	3.5 ± 0.5	0.022 ± 0.005 (24)	2.3 ± 0.5	0.018 ± 0.002 (66)	2.3 ± 0.3
B–D	0.032 ± 0.004 (50)	3.3 ± 0.5	0.022 ± 0.005 (24)	2.3 ± 0.5	0.019 ± 0.003 (67)	2.4 ± 0.3
Ad–At	0.0019 ± 0.0010 (3)	0.20 ± 0.11	ND	ND	0.0044 ± 0.0011 (16)	0.56 ± 0.14
Ad–Ah	0.0012 ± 0.0012 (2)	0.12 ± 0.08	ND	ND	0.0061 ± 0.0013 (22)	0.78 ± 0.17
At–Ah	0.0019 ± 0.0010 (3)	0.20 ± 0.11	ND	ND	0.0044 ± 0.0012 (16)	0.56 ± 0.15
Bt–Bh	0.0000 ± 0.0000 (0)	0.00 ± 0.00	ND	ND	0.0000 ± 0.0000 (0)	0.00 ± 0.00
Dd–Dh	0.0037 ± 0.0017 (6)	0.39 ± 0.16	ND	ND	0.0038 ± 0.0011 (14)	0.49 ± 0.14
Acc-2						
Pooideae–Panicoideae	0.578 ± 0.029	60	ND	ND	n.a.	n.a.
Rice–*Panicoideae*	0.570 ± 0.029	60	ND	ND	n.a.	n.a.
Rice–*Pooideae*	0.534 ± 0.028	55.4 ± 2.9	ND	ND	n.a.	n.a.
Wheat–*Brachypodium*	0.359 ± 0.017	37.3 ± 1.8	ND	ND	n.a.	n.a.
Wheat–barley	0.100 ± 0.009 (132)	10.6 ± 0.9	ND	ND	0.076 ± 0.005 (209)	10.6
A–B	0.084 ± 0.009 (126)	8.7 ± 0.9	ND	ND	0.055 ± 0.004 (153)	7.7 ± 0.6
A–D	0.074 ± 0.007 (112)	7.7 ± 0.8	ND	ND	0.048 ± 0.004 (135)	6.7 ± 0.6
B–D	0.049 ± 0.006 (76)	5.1 ± 0.6	ND	ND	0.034 ± 0.004 (97)	4.8 ± 0.5
Ad–Ah	0.0076 ± 0.0022 (12)	0.79 ± 0.23	ND	ND	0.0052 ± 0.0014 (15)	0.73 ± 0.20
Bt–Bh	0.0025 ± 0.0012 (4)	0.26 ± 0.13	ND	ND	0.0017 ± 0.0008 (5)	0.24 ± 0.11
Dd–Dh	0.0006 ± 0.0006 (1)	0.06 ± 0.06	ND	ND	0.0007 ± 0.0005 (2)	0.10 ± 0.07

The molecular clock was calibrated using 60 MYA (underlined) for divergence of *Pooideae* from *Panicoideae*. The molecular clock for intronic sequences was calibrated using dates for divergence of wheat and barley (underlined) calculated from synonymous substitutions in each gene. Ea fragment encodes amino acids 1–1498. Number of nucleotide substitutions is shown in parentheses. The first intron (in 5′-UTR) and the last intron (in the CT domain) were excluded from the comparisons. Average number of synonymous sites in *Acc-1* comparisons: 1,610 (full-length), 1,100 (Ea segment), 1,358 (partial sequence for comparisons with barley), 845 (partial sequence of the Ea segment for comparisons with barley). Average number of synonymous sites in *Acc-2* comparisons: 1,583 (full-length), 1,404 (partial sequence for comparisons with barley). Number of intron sites in *Acc-1* and *Acc-2* comparisons, 3,651 and 2,893, respectively. A, B, and D indicate *Acc* homoeologs from diploid (d) *T. urartu* or *Ae. tauschii*, tetraploid (t) *T. turgidum*, and hexaploid (h) *T. aestivum*. Errors inherent in the sequence datasets were calculated by the bootstrap method. n.a., not available; ND, not determined.

Fig. 3*E*). Average synonymous and intron rates are two to three times higher in the *Acc-2* than in *Acc-1* genes. Significantly more homoeolog-specific substitutions accumulated in the A and B *Acc-2* genes. Uneven distribution of the synonymous substitutions along the gene is most striking for the A to B comparison (Fig. 2*E*).

The *Acc-2* B locus in *T. turgidum* and *T. aestivum* is characterized by a significant divergence upstream of *Acc-2* and low divergence downstream of the gene. This observation might be explained by haplotype variation coupled with a recombination event whose cross-over was located downstream of *Acc-2* (Fig. 3*F* and Fig. S1*B*). For the *Acc-2* A locus in *T. turgidum* and *T. aestivum*, only the upstream flanking region was available for comparison (Fig. 3*G*). The *Acc-2* D locus shows the same correlation as the B locus, except that in this case, the rate difference between the upstream and downstream flanking regions is only 2-fold (Fig. 3*H*). With the exception of the downstream region at the B locus, the nucleotide substitution rates in the intergenic region are 2- to 9-fold higher than either intron or synonymous rates within the *Acc-2* gene. Similar to the *Acc-1* locus, these rates vary significantly across the locus (Fig. S1*B*).

Pairwise distances at synonymous sites and intron distances are shown in Table 1 to describe quantitatively some of the observations presented above and to estimate divergence times. The molecular clock was set by using 60 Mya for divergence of *Pooideae* and *Panicoideae*. The molecular clock for introns, which can be aligned reliably only at the tribe level or below, was calibrated by using dates for divergence of wheat and barley calculated from synonymous substitutions in each gene (Table 1). Two distances were calculated for the *Acc-1* gene, one based on the full-length sequence and one based on the first two-thirds (segment Ea) of the coding sequence. The two distances are different for A, B, and D homoeologs and between wheat and barley but are identical (within error) for all other sequence pairs. This result is consistent with increased incidence of substitutions in the terminal third of the *Acc-1* coding sequence (segment Eb) observed only at the homoeolog level and affecting significantly only the distances between pairs of A, B, and D genes and, to a lesser degree, distances between barley and wheat A, B, and D sequences. This phenomenon does not appear to be a general property of the *Acc-1* gene. The sequence of segment Ea was also used to compare distances at synonymous and intron sites. Intron distances are very similar to distances at synonymous sites in neighboring exons (Fig. 3). The divergence times calculated by using the two data sets are consistent (Table 1).

Discussion

Our analysis indicates a departure from neutral evolution at both *Acc* loci; substitution rates at synonymous and intron sites vary along the *Acc* genes (Fig. 2) and, on average, these rates are lower than substitution rates outside of the coding part of the genes (Figs.

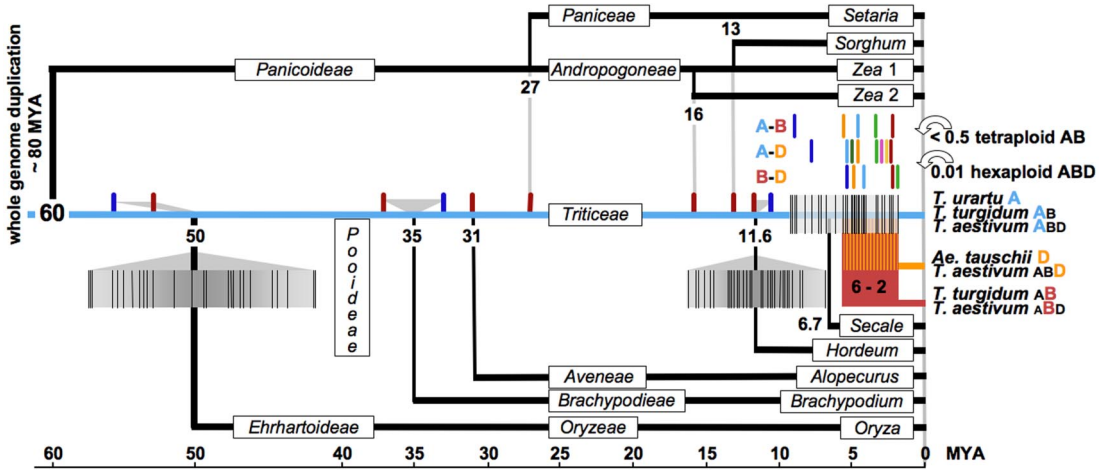

Fig. 4. Time line of wheat evolution. Divergence times were calculated by using 60 Mya for the divergence of *Pooideae* from *Panicoideae* and 11.6 Mya for the divergence of barley from wheat to calibrate the molecular clock. The latter calibration point was used only for *Triticum* and *Aegilops* genes. Each thin black vertical line in areas shaded in gray indicates a single estimate of divergence time for rice and wheat, barley and wheat, and between *Triticum* and *Aegilops* genes (Table S2 and Table S3). Colored marks above the blue line indicate divergence dates estimated based on *Acc-1* (red) and *Acc-2* (dark blue) genes. Divergence times between A and B, A and D, and B and D homoeologs are shown in rows marked by the letters. These estimates are based on the *Acc-1* and *Acc-2* genes, as well as the *Pgk-1* gene (light green), the *Phy*C gene (orange), the *Sut*1 gene (light blue), the Q gene (dark green), the 6GPDH gene (violet), and an average based on four genes reported previously (yellow). Numbers indicate date estimates for different nodes. The upper estimate for the formation of tetraploid wheat is based on the *Acc-1* sequence. The time of formation of the hexaploid wheat genome is based on archeological artifacts.

2 and 3, Fig. S1). The significance of these differences is difficult to gauge because evolution of noncoding gene sequences as well as intergenic sequences is poorly understood. Furthermore, comparisons among homoeologous loci, the only level at which intergenic sequences can be compared, may be affected by haplotype variation. On the other hand, substitution rates at synonymous and intron sites in the same part of the gene are very similar (Figs. 2 and 3, Table 1). The number of synonymous substitutions among homoeologs is low for a reliable estimate, even for genes as large as *Acc*. Even then, synonymous and intron rates are very similar, except for A homoeologs of *Acc-1* (Table 1, Fig. 3*C*).

The promoter regions of the *Acc* genes are very compact: no more than 1 kb, Fig. 2*A*. This is evident from sequence comparisons of the homoeologous genes. In the case of the A homoeologs of *Acc-1* and *Acc-2*, the upstream sequences were changed completely by large indels within a short distance of the translation start codon (Fig. 1). Large blocks of conserved sequence are present at 5′- and 3′- ends of both *Acc-1* and *Acc-2* when A, B, and D homoeologs are compared, but only *Acc-1* sequences could be used in rate calculations (Figs. 2*C* and 3*A*). Such comparisons for the *Acc-2* genes would require significant manual editing of sequence alignments. Substitution rate in the noncoding 5′- and 3′-ends of *Acc-1* is on average 3.5-fold higher than either synonymous or intron rate.

Comparisons of intergenic sequences at the *Acc* loci were only possible for pairs of homoeologs (Fig. S1); there is no sequence similarity between A, B, and D homoeologs outside of the genes shown in Fig. 1. Comparison of gene and intergenic sequences poses another problem; the intergenic rates vary significantly across the locus (Fig. S1), reflecting, most likely, a combination of DNA composition (active genes, various repetitive sequences, DNA methylation) as well as haplotype variation. There are few or no genes other than *Acc* predicted for the loci (Fig. 1 and gene models shown in the annotation files at www.tigr.org/tdb/e2k1/tae1). The average GC content of regions flanking *Acc* genes is 45% compared with 33% in introns. The analysis is further complicated by low information content for synonymous sites and absence of introns from the 3′-end of the *Acc* coding sequence. Haplotype variation and recombination, such as the postulated recombinations with crossover sites located within or near the *Acc-2* B and D genes (above), could lead to either over- or underestimation of relative rates. The average substitution rate in intergenic regions is 3.5 times higher than at intron sites, based on the data shown in Fig. 3 *C*, *D*, and *G*.

Assuming the intron substitution rate in *Acc* genes in *Triticeae* to be 3.6×10^{-9} substitutions per site per year (average intron substitution rate between barley and wheat is 0.084 Table 1; and divergence time is 11.6 Mya, Fig. 4), the average substitution rate in the intergenic region calculated as the number of substitutions per site per year is 13×10^{-9}, similar to the rate reported previously (18). Clearly, synonymous and intron substitution rates do not directly reveal nucleotide substitution rates in the flanking intergenic regions.

The average distance at synonymous sites for full-length *Acc-1* is 0.58, 0.56, and 0.50 when *Pooideae* and *Panicoideae*, rice and *Panicoideae*, and rice and *Pooideae* sequences, respectively, are compared (Table 1). For *Acc-2*, the corresponding distances are 0.58, 0.57, and 0.53. These numbers are very similar to the multigene averages of 0.59, 0.56, and 0.48, respectively, calculated by us (Table S2) as well as distances previously determined from genomic surveys of multiple maize, rice, sorghum, and barley genes (9, 11). The corresponding multigene average substitution rate (including *Acc* genes) is 4.8×10^{-9} substitutions per synonymous site per year, assuming 60 Mya for the time *Pooideae* (wheat, barley) and *Ehrhartoideae* (rice) diverged from *Panicoideae* (maize, sorghum). This rate is lower than a commonly used previous estimate of 6.0×10^{-9} (19), but the lower estimate for the time of divergence of maize and rice (50 Mya) used in that study to calibrate the molecular clock accounts for most of the difference.

The 50 Mya time of divergence of rice from *Triticeae* estimated from sequences of multiple genes (Fig. 4) is consistent with the prevailing belief that wheat and rice diverged from each other only after they diverged from maize. Full-length *Acc* sequences were also used to estimate the divergence time of *Aveneae* (*Alopecurus*) from *Triticeae* at 31 Mya, *Brachypodieae* (*Brachypodium*) from *Triticeae* at 35 Mya, and *Paniceae* (*Setaria*) from *Andropogoneae* (maize) at 27 Mya (Table 1). The *Brachypodium*–*Triticeae* divergence time is within the range of a previous multigene estimate [35–40 Mya (20)]. The *Setaria*–maize divergence time is higher than the 24 Mya obtained previously from partial *Acc-1* sequences of *Panicum virgatum* and maize (21). Finally, the divergence time of the two copies of maize *Acc-1* gene at 16 Mya and the divergence time of maize and *Sorghum* at 13 Mya are consistent with the previously

EVOLUTION

proposed scenario of evolution of these species including a whole genome duplication (polyploidization) of the maize genome (10).

Barley is an important outgroup for molecular evolution studies of the *Triticum/Aegilops* complex. The use of the divergence time between wheat and barley to set the molecular clock to date the subsequent events in wheat evolution may help minimize effects of substitution rate variability over longer distances and in different less closely related lineages and should avoid problems with the clock calibration for individual genes caused by differential loss of duplicated genes. In addition, intronic sequences can often be aligned with high confidence and used for the evolutionary analysis of the *Triticeae* species. For *Acc* genes, intron and synonymous substitution rates produce consistent estimates of divergence times when neighboring introns and exons are compared, but introns have a much higher information content (Fig. 3 and Table 1).

The distance between wheat and barley calculated for synonymous sites is 0.107 for *Acc-1* (Ea segment) and 0.100 for *Acc-2* (full-length), and the corresponding divergence times are 11.8 and 10.6 Mya (Table 1). These numbers are very similar to the multigene average estimates of 0.114 and 11.6 Mya, respectively (Fig. 4 and Table S3). The wheat–barley divergence time estimates vary between 6.7 and 16.2 Mya, reflecting properties and evolutionary history of individual genes.

By using 11.6 Mya for the time of wheat divergence from barley and coding and intron sequences of the first two-thirds of the *Acc-1* genes, the three divergence times calculated for the A, B, and D homoeologous genomes are 2.3–2.4 Mya, consistent with previous estimates based on a short fragment of the gene and by using two alternative evolutionary models (4, 14). Note that for the *Acc-1* gene, these estimates do not differ from the estimates using a much more distant event to calibrate the clock (Table 1).

We analyzed a group of genes for which barley, rice, maize, and multiple wheat sequences, putative homoeologous sequences, are available (Fig. 4 and Tables S2 and Table S3). Divergence times between these wheat genes, calculated relative to the divergence of the same barley gene, vary between 2 and 9 Mya. Only several of these genes have been studied more extensively, allowing phylogenetic analysis. The divergence times for these genes vary between 2 and 6 Mya, with the exception of A-B and A-D divergence of *Acc-2* at 9 and 8 Mya, respectively. We argue that the estimates at the lower end of the spectrum (2–3 Mya) reveal more accurately the divergence time of the ancestors of the A, B, and D genomes of wheat and its relatives. The higher estimates probably capture haplotype variation, gene duplication, and specific loss, including the possibility of formation of chimeric genes by unequal crossing over between duplicated genes.

The large size of the *Acc* genes allowed us to identify sequence variations of unknown origin affecting divergence time estimates: the increased incidence of synonymous substitutions in the last third of the *Acc-1* gene and a significant variation of the substitution rate along the *Acc-2* gene. In the former case, the extra substitutions increase the divergence time between A, B, and D homoeologs by only ≈0.5 Mya, if a proper barley reference sequence is used for the molecular clock calibration. In the latter case, the divergence time estimates could differ by as much as 4-fold, if 1-kb coding sequences from different parts of the gene are compared. Different haplotypes or gene copies are a possible explanation of the observed sequence variation. Haplotype divergence for A, B, and D genomes is significant and can reach 1 million years at some loci (refs. 1, 4, 5, 13, and 22 and this work). The haplotype divergence observed for *A. speltoides*, the progenitor of the B genome, is even greater and can exceed the divergence between A, B, and D homoeologs (23). Both *Acc-1* and *Acc-2* genes were duplicated at least once after divergence of *Triticum/Aegilops* from barley (1, 5) followed by specific loss of one gene copy in different species.

The composition of the *Acc* homoeoloci suggests a significant haplotype divergence, which in some cases reaches or even exceeds 0.5 million years (Table 1). Such variation might result in a significant overestimate of the time of tetraploid wheat formation, which, based on the *Acc-1* estimates shown in Table 1, occurred no more than ≈0.5 Mya. This upper estimate is consistent with results published previously (4, 13). The date was deduced from the comparison of the *Acc-1* A homoeolog in *T. urartu* to *T. turgidum* and *T. aestivum* (average of two synonymous and two intron estimates). The estimates based on synonymous substitutions are only approximate due to very low sequence divergence. The estimate based on the intronic sequences, which have a higher information content, is slightly higher: ≈0.7 Mya.

Materials and Methods

Details of the BAC sequencing strategy can be found in the *SI Text*, along with the methods used for assembly and alignment. Substitution rates and divergence times were calculated by using standard methods. Additional references are given in the *SI Text*.

ACKNOWLEDGMENTS. We thank A. R. Akhunova and J. Dvorak (University of California, Davis) and W. Li and B. Gill (Kansas State University, Manhattan, KS) for providing BACs; J. Carroll and W. J. Buikema (University of Chicago CRC DNA Sequencing Facility, Chicago) for BAC sequencing; and W. Zhu (The Institute for Genomic Research, Rockville, MD) for annotation of BAC sequences. The hexaploid wheat BAC library is a resource obtained by the Genoplante consortium (www.genoplante.com). This project was supported in part by a grant from U.S. Department of Agriculture, National Research Initiative Competitive Grants Program.

1. Huang SX, *et al.* (2002) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol Biol* 48:805–820.
2. Zuther E, *et al.* (2004) Complex nested promoters control tissue-specific expression of acetyl-CoA carboxylase genes in wheat. *Proc Natl Acad Sci USA* 101:1403–1408.
3. Podkowinski J, *et al.* (2003) Expression of cytosolic and plastid acetyl-CoA carboxylase genes in young wheat plants. *Plant Physiol* 131:763–772.
4. Huang S, *et al.* (2002) Genes encoding plastid acetyl-CoA carboxylase and 3- phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci USA* 99:8133–8138.
5. Faris J, Sirikhachornkit A, Haselkorn R, Gill B, Gornicki P (2001) Chromosome mapping and phylogenetic analysis of the cytosolic acetyl-CoA carboxylase loci in wheat. *Mol Biol Evol* 18:1720–1733.
6. Gornicki P, *et al.* (1997) Plastid-localized acetyl-CoA carboxylase of bread wheat is encoded by a single gene on each of the three ancestral chromosome sets. *Proc Natl Acad Sci USA* 94:14179–14185.
7. Podkowinski J, Sroga GE, Haselkorn R, Gornicki P (1996) Structure of a gene encoding a cytosolic acetyl-CoA carboxylase of hexaploid wheat. *Proc Natl Acad Sci USA* 93:1870–1874.
8. Gornicki P, *et al.* (1994) Wheat acetyl-CoA carboxylase: cDNA and protein structure. *Proc Natl Acad Sci USA* 91:6860–6864.
9. Wang X, Shi X, Hao B, Ge S, Luo J (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol* 165:937–946.
10. Wei F, *et al.* (2007) Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. *PLoS Genet* 3:e123.
11. Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 101:9903–9908.
12. Nesbitt M, Samuel D (1996) in *Proceedings of the First International Workshop on Hulled Wheats*, eds Padulosi S, Hammer K, Heller J (International Plant Genetic Resources Institute, Rome), pp 41–100.
13. Dvorak J, Akhunov ED (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops–Triticum alliance. *Genetics* 171:323–332.
14. Devos KM, Beales J, Ogihara Y, Doust AN (2005) Comparative sequence analysis of the phytochrome C gene and its upstream region in allohexaploid wheat reveals new data on the evolution of its three constituent genomes. *Plant Mol Biol* 58:625–641.
15. Crepet WL, Feldman GD (1991) The earliest remains of grasses in the fossil record. *Am J Bot* 78:1010–1014.
16. Prasad V, Stromberg CA, Alimohammadian H, Sahni A (2005) Dinosaur coprolites and the early evolution of grasses and grazers. *Science* 310:1177–1180.
17. Linder HP (1986) The evolutionary history of the Poales/Restionales—A hypothesis. *Kew Bull* 42:297–318.
18. Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol Biol Evol* 23:1386–1396.
19. Gaut BS (1998) Molecular clocks and nucleotide substitution rates in higher plants. *Evol Biol* 30:93–120.
20. Bossolini E, Wicker T, Knobel PA, Keller B (2007) Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. *Plant J* 49:704–717.
21. Huang S, Su X, Haselkorn R, Gornicki P (2003) Evolution of switchgrass (*Panicum virgatum* L.) based on the sequence of the nuclear gene encoding plastid acetyl-CoA carboxylase. *Plant Sci* 164:43–49.
22. Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866.
23. Kilian B, *et al.* (2007) Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes. *Mol Biol Evol* 24:217–227.